REPORTING OF BLIND METHODS: AN INTERDISCIPLINARY SURVEY

by CAROLINE WATT and MARLEEN NAGTEGAAL

ABSTRACT

The use of blind methods can help minimise experimenter bias and therefore can be one indicator of methodological quality. We present an interdisciplinary survey of the reporting of blind methods in scientific journals. The survey, of 1214 papers, aimed to replicate and extend upon an earlier survey. The findings showed very high inter-coder reliability and confirmed the overall pattern found in the previous survey, with parapsychology showing the highest level of reporting of blind methods (79.1%), and the lowest level being found for the physical sciences (0.5%). The implications of these findings are considered, and difficulties in making methodological comparisons across diverse research paradigms are highlighted.

INTRODUCTION

The goal of science is to advance understanding. However, research findings may be rendered invalid or unreliable in a number of ways. Barber (1976) lists ten possible pitfalls, including the 'Investigator Data Analysis Effect', the 'Investigator Loose Procedure Effect', and the 'Experimenter Unintentional Expectancy Effect'. The present survey focuses on the latter experimenter expectancy effects, whereby the experimenter may subtly influence a participant to respond so as to confirm the experimenter's expectancies or desires (e.g. Rosenthal, 1976; Rosenthal & Rubin, 1978). Such expectancy effects may not be limited to research with human participants. For example, it is possible that even in research with inanimate systems such as plants or chemical samples, differential handling of samples, or errors in data-recording, may allow the experimenter's expectancies to affect the research results.

One way in which the possibility of such inadvertent experimenter bias may be minimised is to use 'blind' methods, whereby the experimenter is kept unaware of potentially important aspects of the participants or the task. For example, in medical research testing the effects of drug A compared with placebo B, the experimenter is said to be blind if he or she does not know whether the patient has been given A or B. If neither the experimenter *nor* the patient knows whether A or B has been administered, the design is said to be double-blind (though if the patient begins to experience side-effects of the drug, the condition allocation may be inadvertently revealed). The assessment of the patient is therefore not biased by the experimenter's or the patient's expectations of the likely effects of A or B. In the case of research with inanimate systems, such as a test-tube containing a blood sample, we presume that the system under study is already 'blind' in that it has no expectancies. However, bias in handling of materials, data recording, and so on, may still be reduced by keeping the experimenter blind to key aspects of the sample or task, such as whether the sample is from a patient who received drug A or from a patient who received placebo B.

In parapsychology, particularly in ESP research, blind methods are most often found when efforts are made to conceal the ESP target identity. It is trivially true that the telepathic 'receiver' is blind, because that is the nature of the psi task. However, as methodologies have tightened up over the years, it is now fairly common practice also to keep others involved in the experiment blind as to the target identity, notably the experimenter or the person who is recording the receiver's impressions. Sometimes the target is chosen by computer and is known only to the 'sender'; at other times, a person is given the task of randomising and producing a concealed target order. That person should thereafter have no contact with those involved in the experiment, in order to try to avoid leakage of target information and breaking the blind. We will come back to the issue of how blinds are typically operationalised in parapsychology, compared with other disciplines, in the Discussion section.

Blind methods can reduce bias from experimenter expectancy effects; therefore the use of blind methods can be an indicator of methodological quality (though the study could still be of poor quality in other aspects of its methodology). To investigate the attention paid to possible experimenter effects across different disciplines, Dr Rupert Sheldrake (RS) conducted a survey of the reporting of blind methods in leading scientific journals (Sheldrake, 1999). This survey found that blind methods were rarely reported in the 'hard' sciences of physics, chemistry and biology. In the human and behavioural sciences the most widespread reporting of blind methods was in parapsychology, for which 85% of applicable papers reported blind methods, compared with only 24% in medical sciences (Sheldrake, 1999).

The present paper reports a conceptual replication and extension of RS's survey. This was achieved by adding independent duplicate coding of articles, by surveying a more recent sample of the same journals as were surveyed by RS, and by doubling the number of parapsychology journals surveyed.

In the original survey, a single person (a research assistant) coded the majority of the papers for whether or not it was appropriate to include them in the survey (for example, literature reviews and theoretical papers were not included), and for whether or not they reported blind methods. If there was any uncertainty about a coding, the paper in question was referred to RS for a decision (Sheldrake, personal communication with CW, 19th April 1999). It appears that there was no independent (duplicate) coding, so that one cannot assess the reliability of the coding judgements in the original survey. Our replication survey therefore introduced this methodological improvement. Additionally, the original survey covered just two parapsychology journals, *Journal of Parapsychology (JP)* and the *Journal of the Society for Psychical Research (JSPR),* which yielded only 27 applicable papers. This was a small sample compared with the hundreds of papers surveyed from other fields. In order to extend this to include the other two main English-language parapsychology journals active at that time, and to increase the number of applicable papers from parapsychology our replication survey also covered the *European Journal of Parapsychology (EJP)* and the *Journal of the American Society for Psychical Research (JASPR).*

METHOD

Inclusion Criteria

In terms of the specific journal issues included, our survey did not overlap at all with RS's, so it does not attempt to replicate his specific findings. We aimed to conduct a conceptual replication, with an expanded and updated set of journals, so we examined the most recently available issues of the mainstream journals used in the previous survey (Sheldrake, 1999).[1] Because our survey was conducted in mid-1999, the majority of journals surveyed were published in 1999. To maximise the number of parapsychology articles surveyed, *JP* and *JSPR* coverage picked up from where the previous survey left off (from 1997), *EJP* was surveyed from 1990–1999,[2] and *JASPR* was surveyed from 1995 to 1998 (due to publication delays, the later *JASPRs* were not yet available). See Table 1 for details of specific volume numbers surveyed. We attempted to match the number of mainstream papers reviewed by RS by surveying the same number of issues for each journal as he had. For example, RS surveyed three consecutive parts of the *Journal of the American Chemical Society* (Volume 118, Parts 39–41), so we surveyed the three most recently available consecutive parts of the same journal (Volume 121, Parts 25–27).

Coding Criteria

We attempted to use the same coding criteria as were used in the original survey. There were two criteria for coding. Firstly, each article was categorised as to whether it reported an experimental study; if so, it was judged applicable for inclusion in the survey. Theoretical, review and survey papers were excluded. Secondly, applicable papers were carefully read, focusing particularly on the Methods sections, and a judgement was made as to whether or not the study reported blind methods (single-blind or double-blind). This judgement was based on what was said in the paper about the study's methodology and procedure. Sometimes it was obvious that the paper was reporting blind methods; for instance, the words 'blind' or 'double-blind' might be used, or there might be a description of how allocation to experimental and control conditions was concealed from the experimenters, or it might be said that the experimenter was unaware of the identity of the sample that he or she was observing or measuring. Less often, the use of blind methods was inferred from the procedural description. For example, although there was no discussion of how the condition allocations were concealed, there might be a line in the procedure describing how these were revealed at the end of the data collection or analysis (hence, by implication, they must first have been concealed). To give a concrete example of a paper that was coded as 'blind', one of the biology papers from the journal, *Nature,* (Keller & Ross, 1998) described genetic

[1] Sheldrake's survey included journals from 1996 to 1998, with the exception of the parapsychology journals, which dated from 1993 to 1996. We arbitrarily decided not to include the *Proceedings of the National Academy of Sciences (US),* because we felt that the biological sciences were already well represented by the other seven biological sciences journals. Even excluding this journal, the biological sciences still represented the majority of the papers surveyed.

[2] *EJP* published only one issue per year; therefore the survey started in 1990 in order to retrieve similar numbers of articles to the more frequently published parapsychology journals.

influences on aggression in red fire ants. In two places blind methods were mentioned: "Assessment of the level of aggression was done without knowledge of *Gp–9* genotypes" (p.574); and "Scoring was done without knowledge of whether test workers had been rubbed against attacked (*BB*) or non-attacked (*Bb*) queens" (p.575).

It is possible that in some cases blind methods were being used but not reported. Given that it usually takes a certain amount of planning and effort to use blind methods in a study, we believe it would be unusual not to mention this in the write-up. However, we cannot be certain to have included papers that used blind methods but did not report them: that is why this paper refers to the 'reporting' of blind methods. In this we have been more conservative than RS, whose survey referred to the 'use' of blind methods (Sheldrake, 1999).

The criterion for whether or not a study was coded as blind was quite liberal, in that a paper was coded as reporting blind methods if such methods were mentioned at *any* point during the experimental protocol, even if blind methods were not included at every possible point during the protocol. So far as we could determine, these were the same criteria used by Sheldrake, and the intention was basically to ascertain whether an investigator was showing some awareness of blind methods, even if he or she was apparently not using them to the greatest possible extent. Using these criteria, 1214 applicable papers were identified.

Duplicate Coding

The coding was done by Watt and Nagtegaal, including independent blind duplicate coding on 22% of the journals. That is, Watt coded approximately half of the journals, Nagtegaal coded approximately half of the journals, and Watt coded a portion of the journals that had also been coded by Nagtegaal, unaware of Nagtegaal's coding, and vice versa. The double-coded journals were two of the parapsychology journals, and the three psychology journals. These were chosen on the basis of RS's survey, which showed that blind methods were being reported in these journals, whereas blind methods appeared to be extremely rare (less than one per cent) in the physics, chemistry and biology journals. We therefore judged that if the journals were likely to have a reasonable proportion of blind studies, this would provide a more sensitive indicator of any potential discrepancy between the two coders.

RESULTS

Coding Reliability

Extremely high inter-rater reliability was found, both for the decision on which papers were applicable ($\kappa = 0.96$, N = 150) and for the decision on which papers reported blind methods ($\kappa = 0.90$, N = 76) (Cohen's Kappa indicates the proportion of agreement between raters after chance agreement has been removed — Rosenthal & Rosnow, 1991). This high reliability suggests that our findings are valid in relation to our criteria for applicability and for categorisation of papers.

Reporting of Blind Methods

Table 1 gives details of the journals and volume numbers surveyed, the number of applicable papers, and the number reporting blind methods. For

comparison purposes, the table also gives summary figures from RS's (1999) survey.

Table 1

Journal Volumes (Part Numbers in Brackets) Surveyed, Numbers of Applicable Papers, and Number Reporting Blind Methods

Journal	Applicable papers	Blind methods
Physical Sciences		
Journal of the American Chemical Society Vol. 121 (25–27)	110	0
Journal of Applied Physics Vol. 81 (1)	87	1
Journal of Physics (Condensed Matter) Vol. 11 (18–19)	21	0
Totals	**218**	**1 (0.5%)**
Sheldrake (1999)	237	0
Biological Sciences		
Biochemical Journal Vols. 340–341 (1–3; 1)	133	0
Cell Vol. 97 (6–7)	23	2
Heredity Vols. 81–82 (6; 1–4)	53	2
Journal of Experimental Botany Vols. 49–50 (327–334)	110	1
Journal of Molecular Biology Vols. 289–290 (4; 1–3)	91	1
Journal of Physiology Vols. 514–515 (1–3; 1–3)	151	4
Nature (biology papers) Vols. 393–394 (6686–6696)	100	6
Totals	661	16 (2.4%)
Sheldrake (1999)	914	7 (0.8%)
Medical Sciences		
American Journal of Medicine Vol. 106 (1–5)	34	15
Annals of Internal Medicine Vol. 130 (5–10)	17	8
British Journal of Clinical Pharmacology Vol. 46 (4–6)	30	11
British Medical Journal Vol. 317 (7163–7168)	22	8
New England Journal of Medicine Vol. 340 (16–23)	30	7
Totals	**133**	**49 (36.8%)**
Sheldrake (1999)	227	55 (24.2%)
Psychology and Animal Behaviour		
Animal Behaviour Vol. 56 (1–4)	97	9
British Journal of Psychology Vol. 89 (1–3)	20	3
Journal of Experimental Psychology [JEP]: General Vol. 127 (1–3)	12	2
JEP: Human Perception & Performance Vol. 23 (5–6)	30	9
Totals	**159**	**23 (14.5%)**
Sheldrake (1999)	143	7 (4.9%)

Journal	Applicable papers	Blind methods
Parapsychology		
Journal of the Society for Psychical Research Vols. 61–63 (846–855)	6	3
Journal of Parapsychology Vols. 60–62 (3–4)	11	6
European Journal of Parapsychology Vols. 8–14	19	18
Journal of the American Society for Psychical Research Vols. 89–92 (1–4; 1–4; 1–4; 1–2)	7	7
Totals	**43**	**34 (79.1%)**
Sheldrake (1999)	27	23 (85.2%)

Note: The summary figures for Sheldrake's survey are included for comparison purposes.

The overall pattern of results essentially replicates that found in the survey conducted by RS. The highest frequency of reporting of blind methods was found in parapsychology, and the lowest was found for journals in the physical and biological sciences. The only significant difference between our survey and RS's is that we found a higher proportion of psychology and animal behaviour papers reporting blind methods, compared with the previous survey (14.5% versus 4.9%; $\chi^2 = 4.5$, $p < 0.05$, df = 1). There is no obvious explanation for this discrepancy. For the other disciplines compared, there were no significant differences, suggesting there was no systematic difference between our coding and RS's (physical sciences $\chi^2 = 0.30$, df = 1; biological sciences $\chi^2 = 0.007$, df = 1; medical sciences $\chi^2 = 0.01$, df = 1; parapsychology $\chi^2 = 0.52$, df = 1). Possibly there were simply more blind papers in the psychology and animal behaviour journals included in our survey, compared with those in the previous survey.

Discussion

The present survey aimed to update, replicate and extend upon a previous survey by Sheldrake (1999) of the reporting of blind methods in different scientific disciplines. Our tests of inter-rater reliability suggest that our judgements about paper applicability and reporting of blind methods are reliable. Our pattern of findings essentially replicates that of the previous survey: highest reporting of blind methods was found in parapsychology, while blind methods were rarely reported in the physical and biological sciences. By doubling the number of parapsychology journals surveyed, we also hoped to provide a more representative picture of this field. However, the number of parapsychology papers surveyed was still necessarily small compared with the mainstream disciplines.

Our results show that blind methods are rarely reported in the physical and biological sciences. This may be either because blind methods are not being used, or because they are not considered to be an important aspect of

methodology. Why might this be? Critics (e.g. Barber, 1976, 1978) of the psychological literature on experimenter expectancy effects have suggested that such effects are far less pervasive than is suggested in that literature (e.g. Rosenthal, 1976; Rosenthal & Rubin, 1978). If this is so, perhaps the subtle biases introduced by using non-blind methods are relatively trivial in physical sciences, where the experimental effect is often so large that there is no need for inferential statistics to detect it (Utts, personal communication with CW, 1999). The need for blind methods in the physical sciences may also be reduced where treatment of samples is done by machine, and where observations of data are made automatically without a great deal of human involvement.

Experimenter bias nevertheless remains a theoretical possibility in the physical and biological sciences. The one blind physics paper we found in our survey described how five identical samples were sent to five different laboratories for positron annihilation spectroscopy. Errors in measurement, and handling differences, were suggested as possible contributory factors to the finding of significant inter-laboratory variation in measurement (Goldberg, Knights, Simpson & Coleman, 1999). Similarly, the discovery in plant biology of genes that are intensely upregulated in response to touch (Braam, Antosiewicz & Purugganan, 1994) suggests a mechanism whereby bias may occur through the unintentional differential handling of samples directly affecting growth. Ironically, the discovery of the touch-sensitive plant genes came about inadvertently, as a handling artefact in research on the effects upon these genes of manually applying a hormone to the plant. Findings like these illustrate that it is quite possible for the 'human factor' in the physical and biological sciences to be a source of variation in observations and measurements. The extent to which that variation may be affected by the experimenter's expectancies is an indicator of the extent to which blind methods might provide a valuable methodological improvement. This is an empirical question that so far appears to have received little or no attention in the so-called hard sciences.

Blind methods are important in disciplines such as psychology where there is a labile or complex interaction between the experimenter and the research participant or materials. Therefore it is surprising that in psychology — the discipline that has published most about experimenter effects — the reporting of blind methods was relatively infrequent.

Also, blind methods are important in areas where there is a weak or controversial effect, such as in parapsychology (note that this is not mutually exclusive from the former type of discipline). In these cases, it is vital for researchers to demonstrate that they have eliminated possible factors that may have artifactually produced an experimental effect. This may be one reason why parapsychology was found to have particularly high levels of reporting of blind methods: extraordinary claims require extraordinary evidence. Related to this is the fact that psi is negatively defined—as an effect in the absence of other 'normal' causal mechanisms. In this case, it again becomes important to demonstrate that normal factors have been eliminated.

Apart from these factors, however, there may be another reason why parapsychology appears to have particularly high levels of reporting of blind methods. This concerns the prevailing research paradigm. Many mainstream

experimental studies are asking whether treatment condition A is different in effect from control condition B. It is of course important and desirable to attempt to keep condition allocation blind in this type of research. In this case, a study would be judged as blind if the experimenter did not know which participants/samples were in which condition. In contrast, parapsychology is unusual in that, in a high proportion of studies, the main experimental question is whether the observed scoring rate (e.g. number of times an ESP target is correctly identified or 'hit') differs from the scoring rate theoretically expected by chance. In this research paradigm, it is critical to conceal the 'right answer'; that is, the target identity. Following years of critique, from inside and outside the field, parapsychologists routinely keep the experimenter blind to the target identity. This is just good methodology, and it is very much taken for granted. For instance, Akers's (1984) detailed survey of possible methodological weaknesses in 54 ESP experiments found that in most cases the experimenter was appropriately blind to the target identity or the participants' calls. The issue of whether the experimenter was blind to condition allocation was not examined by Akers, presumably because he did not consider this to be a pivotal methodological issue for parapsychology research.

We would argue, therefore, that one reason for the substantial discrepancy in the reporting of blind methods when parapsychology is compared with other sciences, is that like is not being compared with like.[3] The basic paradigm in ESP research is the attempt to identify a hidden target correctly. This is not directly comparable with a mainstream paper in which treatment condition A is contrasted to control condition B. In the former paradigm, there is an unambiguous 'right answer' (the target identity), which should be concealed from everyone except the sender. In the latter paradigm, while a difference in scoring or performance might be expected between condition A and condition B, it is exceedingly unlikely that the expected performance would be 100% in condition A (equivalent to the target) and 0% in condition B (equivalent to the decoy). So, although it is still important to conceal condition allocation in the latter paradigm, we would argue that knowledge of condition allocation conveys less information to the experimenter than would knowledge of target identity.

There do exist some mainstream topics, such as lie detection, in which there is a 'right answer' that should be concealed, and which are analogous to the situation in parapsychology. However, it is much more common in the mainstream for the basic question to be a comparison between conditions. Future interdisciplinary surveys should therefore make a finer-grained analysis, examining, for instance, the nature of the information to which the experimenter is blind.

It is therefore premature to conclude, as one journalist did in response to RS's original survey, that "Parapsychology . . . makes far more use of rigorous experimental methods than other scientific disciplines" (Matthews, 1998, p.12). As this discussion has argued, the question of reporting of blind methods is a complex one, affected by assumptions about the size of the effect under study

[3] This issue was pointed out to us by an anonymous referee and by Dr Richard Wiseman, in commenting on an earlier version of this paper.

and the relative impact of any possible experimenter bias, by the degree of human involvement in measurement and recording, by definition and implications of the experimental effect, and perhaps most fundamentally, by differences in experimental paradigms. Blinds may be put into place at various different levels in an experiment—from concealing the 'right answer' in the main dependent variable, to concealing the participant's condition allocation in process-oriented research. In order to make a meaningful comparison between diverse disciplines, careful attention must be paid to ensuring that like is being compared with like. However, we feel that our survey, together with Sheldrake's, should at the very least give pause for thought to those, such as the AAAS members surveyed by McClenon (1982), who would suggest that parapsychology is 'methodologically weak'. For their central research question, parapsychologists use blind methods as a matter of course.

There is one last, ironic, point that is worth making in the pages of a journal of psychical research. If parapsychologists provide evidence that individuals can obtain information through extrasensory perception, then blinds could be effectively rendered useless, not only in parapsychology but in any scientific discipline.

ACKNOWLEDGEMENTS

We are very grateful to Professor Robert Morris, Dr Rupert Sheldrake, and Professor Richard Wiseman for their helpful comments on earlier versions of this article. Thanks also to our anonymous referees for their helpful suggestions for improvement.

Department of Psychology
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
7 George Square
Edinburgh EH8 9JZ caroline.watt@ed.ac.uk

REFERENCES

Akers, C. (1984) Methodological criticisms of parapsychology. In Krippner, S. (ed.) *Advances in Parapsychological Research 4.* Jefferson, NC: McFarland.

Barber, T. X. (1976) *Pitfalls in Human Research: Ten Pivotal Points.* New York: Pergamon.

Barber, T. X. (1978) Expecting expectancy effects: biased data analyses and failure to exclude alternative interpretations in experimenter expectancy research. *Behavioral and Brain Sciences 3,* 388–380.

Braam, J., Sistrunk, D. M., Antosiewicz, D. M. and Purugganan, M. M. (1994) Arabidopsis TCH3 encodes an unusual CA2+ binding protein and shows environmentally induced and tissue-specific regulation. *Plant Cell 6,* 1553–1565.

Goldberg, R. D., Knights, A. P., Simpson, P. J. and Coleman, P. G. (1999) Assessment of the normalization procedure used for interlaboratory comparisons of positron beam measurements. *Journal of Applied Physics 86,* 342–345.

Keller, L. and Ross, K. G. (1998) Selfish genes: a green beard in the red fire ant. *Nature 394,* 573–575.

Matthews, R. (1998) Blind prejudice: 'hard' scientists believe they are immune to bias.

New Scientist, 17th January, 12.

McClenon, J. (1982) A survey of elite scientists: their attitudes towards ESP and parapsychology. *JP 46*, 127–152.

Rosenthal, R. (1976) *Experimenter Effects in Behavioral Research (enlarged edition).* New York: Irvington Press.

Rosenthal, R. and Rosnow, R. L. (1991) *Essentials of Behavioral Research: Methods and data analysis (2nd edition).* New York: McGraw-Hill.

Rosenthal, R. and Rubin, D. B. (1978) Interpersonal expectancy effects: the first 345 studies. *Behavioral and Brain Sciences 3*, 377–386.

Sheldrake, R. (1999) How widely is blind assessment used in scientific research? *Alternative Therapies 5*, 88–91.